OPEN ACCESS

University of BRISTOL

Ahmed, S. A., Dogra, D. P., Kar, S., Kim, B. G., Hill, P., & Bhaskar, H. (2017). Localization of region of interest in surveillance scene. *Multimedia Tools and Applications*, *76*(11), 13651-13680. https://doi.org/10.1007/s11042-016-3762-y

Peer reviewed version

Link to published version (if available):
10.1007/s11042-016-3762-y

Link to publication record in Explore Bristol Research
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online viaSpringer at https://link.springer.com/article/10.1007%2Fs11042-016-3762-y. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research
General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

main text Click here to download Manuscript Area_Localization_Surveillance_v1.tex

Click here to view linked References

Noname manuscript No.
(will be inserted by the editor)

Localization of Region of Interest in Surveillance Scene

Sk. Arif Ahmed · Debi Prosad Dogra · Byung-Gyu Kim · Paul Hill · Harish Bhaskar

Received: date / Accepted: date

Abstract In this paper, we present a method for autonomously detecting and extracting region(s)-of-interest (ROI) from surveillance videos using trajectory-based analysis. Our approach, localizes ROI in a stochastic manner using correlated probability density functions that model motion dynamics of multiple moving targets. The motion dynamics model is built by analyzing trajectories of multiple moving targets and associating importance to regions in the scene. The importance of each region is estimated as a function of the total time spent by multiple targets, their instantaneous velocity and direction of movement whilst passing through that region. We systematically validate our model and benchmark our technique against competing baselines through extensive experimentation using public datasets such as CAVIAR, ViSOR, and CUHK as well as a scenario-specific in-house surveillance dataset. Results obtained have demonstrated the superiority of the proposed technique against a few popular existing state-of-the-art techniques.

National Institute of Technology
Durgapur, India
E-mail: arif.1984.in@ieee.org

Indian Institute of Technology
Bhubaneswar, India
E-mail: dpdogra@iitbbs.ac.in

SunMoon University
Republic of Korea
E-mail: bg.kim@mpcl.sunmoon.ac.kr

Bristol Vision Institute, University of Bristol
Bristol, U.K
E-mail: paul.hill@bristol.ac.uk

Khalifa University of Science, Technology and Research (KUSTAR)
Abu Dhabi, U.A.E
E-mail: harish.bhaskar@kustar.ac.ae

Keywords Trajectory Analysis · Scene Segmentation · Scene Understanding · Object Tracking · Movement Pattern Analysis

1 Introduction & Related Work

The rapid deployment of CCTV-based surveillance systems has led to a significant increase in the volume of video data, thus making visual analytic solutions for surveillance a key example of big-data analysis. Such abundance in data availability with diverse variability has motivated researchers to focus into autonomous scene understanding. While it is important to improve upon existing models of target detection, tracking and recognition, however, high level semantic analysis through behavioral understanding cannot be ignored. This is the need of today's society to deal with complex scenarios for real-time situation awareness. Visual analysis has become an integral part of many applications. Behavior [2], activity [16, 25, 31], and semantics [29, 30] analysis, anomalous activity detection [22, 27, 32, 33], visual surveillance [36] and video summarization [14], scene segmentation [20] or interest area localization [24], and video object retrieval [11], Visual attention detection [34], are some of them to name.

One of the foremost steps in scene recognition and understanding is region(s)-of-interest (ROI) detection. A ROI can be considered as a region that encloses semantically homogeneous information held within a cognitive boundary. Existing methods of ROI detection support static image-based as well as video-based analysis. Some of these methods include, low-level human visual models [10], visual attention-based models [21], saliency-based methods [1, 19], Visual Salience [13], Video attention [12,15] etc. ROI detection techniques are often challenged by high variability in monitoring conditions as well as diversities in the targets appearance and pose. Broadly, ROI detection techniques can be categorized into bottom-up feature-based and top-down knowledge-based approaches [8]. Approaches in the former category aim to localize structural features that are invariant to the aforementioned diversities and use them to detect ROI. Popular approaches under this category include, background subtraction guided salient area detection method [24], SIFT based region localization methods [9], HoG based methods [26], and salience guided methods [1]. In contrast, the top-down approaches usually start building the scene model using contextual or scene-level information [20].

Lately, ROI has been proven to be highly correlated with target movements within a scene. In other words, from the point-of-view of visual surveillance, location of an object /target in the scene is perceived as interesting when a large number of targets approach toward some specific areas of a scene. In such cases, trajectories of the targets can be analyzed to detect such visual elements in a scene that usually attract targets. In existing video analytic-based approaches, static objects present in a scene are usually neglected and left-out as a part of the background or detected using global object-specific models. However, it is important to acknowledge that the movements of tar-

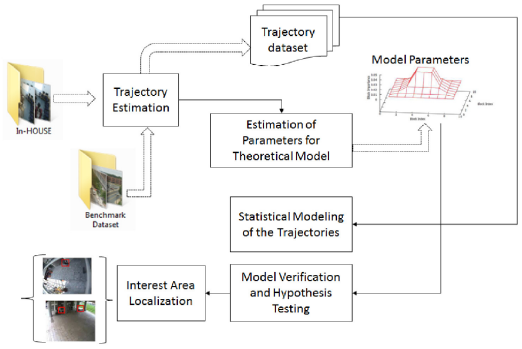


Fig. 1 A functional block diagram of the proposed ROI localization methodology.

gets in unconstrained environment is mainly governed by the presence of such object(s)-of-interest or region(s)-of-interest. By identifying such regions where static objects are located, it is possible to refine the existing decision making processes used in behavior and activity analysis.

In this paper, we have proposed a technique to detect and localize ROI that influence the motion characteristics of other moving targets. A block diagram of the proposed methodology is presented in Figure 1. The rest of the paper is organized as follows. In Section 2, we begin by outlining the main contributions and distinguishing aspects of our work with respect similar techniques available in the literature. A theoretical model of target behavior with a research hypothesis and a novel trajectory analysis technique to validate the hypothesis, are presented in Section 3. In Section 4, we present the experimental validation of our proposed method against baselines. We also present the effect of key system parameters on our proposed model and demonstrate the superiority of the proposed strategy when compared to other baseline techniques. Finally, the research hypothesis has been empirically verified before we conclude in Section 5.

2 Contributions & Distinguishing Aspects

The fundamental purpose of this work is to illustrate a mechanism for ROI localization that allows automatic scene segmentation and thus facilitates making informed decisions on the behavioral understanding of moving targets within a given scene. One key novelty of our method is the integration of behavioral semantics of targets into a theoretical assumption that is based on the distribution of importance of areas using a statistical model of target motion and their interactions. We argue that, the velocity of a target gradually decreases as it approaches toward an object of interest within the scene. It has been observed that, short as well long term analysis are necessary to de-

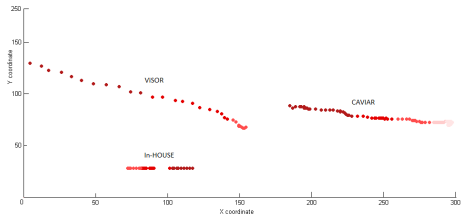


Fig. 2 Examples of real life events (taken from ViSOR, CAVIAR, and In-HOUSE datasets) demonstrating variation in velocity throughout a target's interaction with some object of interest.

tect such velocity changes. Our observations over multiple target trajectories on real-life scenarios has validated the previous claim that a target gradually slows down as it approaches an object of interest within the scene. We present a summary of such observations in Figure 2 on publicly available datasets, where decreasing color saturation highlights the decrease in velocity of moving objects.

In addition, the importance of various regions (represented as blocks) of a surveillance scene is estimated using on an entry-exit model through the correlated measurements of time spent between entry-to-exit, their instantaneous velocity changes, and direction of movements. Preliminary results using the same model without the incorporation of the directional component has already been reported in [4]. Thus the hypothesis of this study focuses on the inclusion of direction for the estimation of the importance of blocks, thereby producing a vector flow field of the target in order to improve the accuracy of localization of the objects of interest and hence the ROI.

3 Proposed Methodology

The method outlined in this research is based on the following underlying principles and assumptions that govern the movements of targets within an unconstrained surveillance environment:

- Targets are allowed to move freely within a surveillance scene and their movements are only restricted by the scene boundary, presence of other moving targets, and presence of static objects of interest.
- Motion dynamics of targets in a scene is mainly influenced by the natural rules of interaction between salient targets and static objects of interest.
- Whilst approaching a static objects of interest or other moving objects, a target usually follows a pattern that can be modeled using simplistic, yet powerful set of primitive features computed from instantaneous velocity and direction of movement.

3.1 Theoretical Model and Problem Formulation

In this subsection, a theoretical model representing the action of a target as it approaches toward a static object of interest, has been introduced. For ease of explanation, let us consider a single static object of interest present in a surveillance scene. Given no restrictions on its movements, a target can access the object of interest from any direction. However, it usually follows the shortest route as depicted in Figure 3.1. Let the scene be divided into rectangular blocks of regions as shown in Figure 3.2 and the possible movements (outer block to inner block) be as given in Figure 3.3-3.6. In such cases, the path possibly includes the following intermediate blocks; C_1, C_2, C_3 assuming the target is initially positioned in one of the outer-layer blocks, e.g. $B_1, ..., B_5$. However, it can reach out to an inner-layer block from any of the immediate outer-layer blocks in various ways. For example, if its initial location is B_1, it must go through block C_1 as depicted in Figure 3.3 considering shortest route. The other possibilities are shown through Figures 3.7-3.8.

Considering the above scenario, a theoretical formulation of the problem can be obtained as described below. Assume that the probability of a target being present inside one of the outer-layer blocks be given as P. According to Figures 3.3-3.6, a target can reach out to one of the inner-layer blocks in three possible ways. Therefore, probability of reaching to any of the inner-layer blocks is three times the probability (i.e. $3P$) of the present outer-layer block. Thus, a target can reach to the object of interest through eight possible ways. Figures 3.7-3.8 explain this assumption considering a three-layer scenario. Without loss of generality, the model can easily be extended to any desired number of layers and the probability values can be computed.

3.2 Research Hypothesis

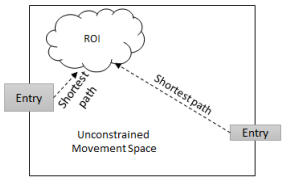
Assuming a surveillance scene is tiled into non-overlapping rectangular blocks, probability of a target τ visiting a block, say b, is denoted by $p_\tau(b) = P_\tau(X = b)$, where $p_\tau(b) \geq 0$, $\forall b$ subject to the condition given in (1)

$$\sum_{b=1}^{N} p_\tau(b) = 1. \tag{1}$$

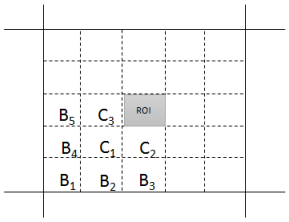
Probability that a particular block, say b, be categorized as interesting is given in (2), where I denotes the importance of b and N represents total number of blocks present in the scene.

$$p(b = ROI) = \max_{i=1}^{N} I(b) \tag{2}$$

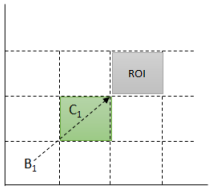
Now, the importance of a block can be estimated using parameters such as time spent and change in instantaneous velocity. However, the resultant direction of a block (θ) also plays an important role. The overall direction of a block



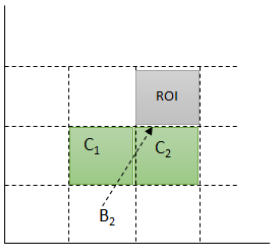
3.1 A typical geometry of the environment with single object of interest.



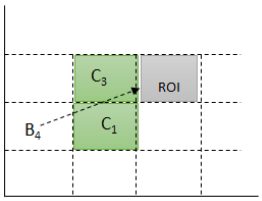
3.2 Division of the area into a rectangular blocks assuming the C_1. object of interest is located at the centre.



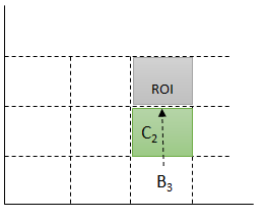
3.3 Movement through



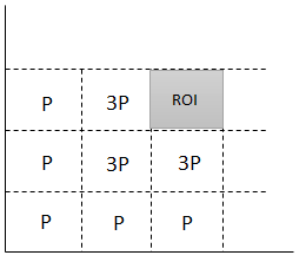
3.4 Movement through C_1 and C_2.



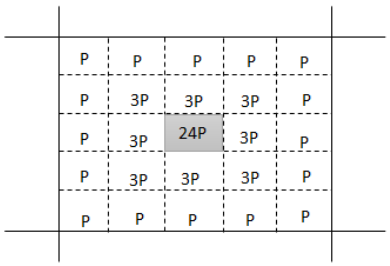
3.5 Movement through C_1 and C_3.



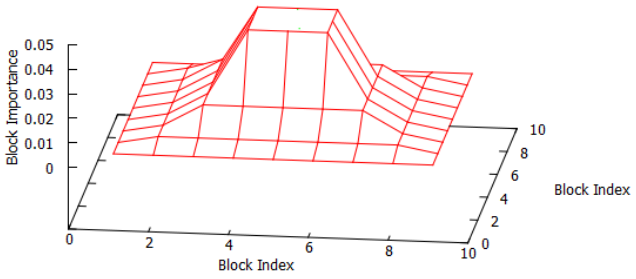
3.6 Movement through C_2.



3.7 Importance of blocks.



3.8 Importance of all surrounding blocks of ROI.



3.9 Representation of access frequency as a pdf of importance.

Fig. 3 Estimation of importance of intermediate blocks around the ROI and theoretical estimation of importance of these blocks in terms of access frequency while the target approaches the region of interest.

represents the total number of dominant directions toward its eight neighbors. Using above three parameters, the importance of a block is formulated as given in (3), where f_1 and f_2 are functions representing linear combinations of these parameters.

$$I(b) \propto f_1(g_b^\tau, v_b^\tau) \star f_2(\theta_b^\tau) \tag{3}$$

Therefore, given a set of trajectories representing target movements inside a surveillance scene, we hypothesize that:

- The correlation coefficients between motion dynamics features including a) the total time spent, b) instantaneous velocity, and c) degree of the block, are indicative of the importance of a given block. The more the value, the more the importance of that block.
- Stochastic modeling of such a system can be approximated using the theoretical assumption shown in Figure 3.9.

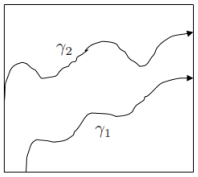
3.3 Feature Extraction

In this section, we describe the method of ROI localization. In order to facilitate the extraction of useful features, we first extract target trajectories using the method proposed in [3]. However, it has been observed that, raw trajectories often contain errors due to varying illumination conditions, occlusions, etc. In order to remove such outlier points from the trajectory set, a simple but effective heuristic has been adopted. A point on the trajectory is removed when a significant deviation is observed from its usual path. This is done as follows. Let, $p(x_i, y_i)$ and $p(x_{i+1}, y_{i+1})$ represent the successive locations of a point on the trajectory at time t_i and t_{i+1}, respectively. We remove the location $p(x_{i+1}, y_{i+1})$ from the trajectory if $\sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} > T$, where T is a threshold that can be estimated empirically. Next, a set of key features are extracted from these refined trajectories to estimate the importance of a block. To begin with, we divide the scene into rectangular blocks of uniform dimension as shown in Figure 3(b). Assume that T represents the set of ℓ trajectories available for training, e.g. $T = \{\gamma_1, \gamma_2 ..., \gamma_\ell\}$, where (4) denotes a trajectory of length $|m_j|$.

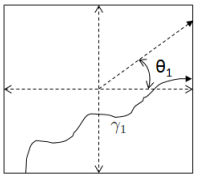
$$\gamma_j = [(x_1, y_1), (x_2, y_2), ..., (x_{m_j}, y_{m_j})] \tag{4}$$

We extract three primitive features namely, a) total time spent by the targets inside a block (global visit count), b) instantaneous velocity of the targets (average instantaneous velocity), and c) direction of movement (degree of a block) whilst passing through a region. These parameters are used in combination to estimate the importance of a block or region.

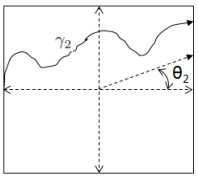
Global Visit Count: Number of times a block (b) is visited by various targets is an important parameter in the present context. We refer this feature as the global visit count (g_b). Initially, the value is set to zero and subsequently



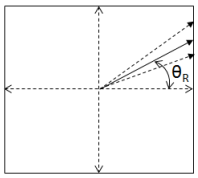
4.1 Examples of two trajectory segments, e.g. t_1 and t_2 entering and exiting a block.



4.2 Direction of the segment t_1 inside the block.



4.3 Direction of the segment t_2 inside the block.



4.4 Dominant direction of the block with respect to t_1 and t_2.

Fig. 4 Estimation of dominant direction of an individual block with respect to its neighborhood.

its value is increased each time a target enters into b. A final estimate of g_b is available once all trajectories are processed.

Average Instantaneous Velocity: It is one of our fundamental assumptions that, instantaneous velocity decreases as the target approaches an object of interest. Therefore, average instantaneous velocity ($\overline{v}^{\tau}$) is computed. To accomplish this, we first computed the minimum v_{min} and maximum v_{max} of the velocity using (5) and (6), where p_l and p_{l+1} denote successive points on a trajectory γ_j bounded by $0 < l < |\gamma_j|$.

$$v^{\tau}_{min} = \min |p_l - p_{l+1}| \tag{5}$$

$$v^{\tau}_{max} = \max |p_l - p_{l+1}| \tag{6}$$

In the next step, $[v^{\tau}_{max} - v^{\tau}_{min}]$ is divided into uniformly spaced non-overlapping segments of equal length and a histogram of instantaneous velocity is generated. Finally, the largest bin is used to estimate the average instantaneous velocity. This mechanism successfully removes any bias that may be induced due to the presence of fast moving segments in a trajectory.

Degree of a Block: To estimate the degree (both out-degree and in-degree inclusive), first we compute the dominant direction $\hat{\theta}$ of a block. Dominant

direction is found using the method depicted in Figure 3.3. Assume targets have accessed a specific block, say b, through several segments. Let, a trajectory segment (s_j) be represented with a set of spatial locations e.g. $[(x_1, y_1), (x_2, y_2), (x_3, y_3), ...(x_z, y_z)]$. The direction of the segment with respect to the block's centre is considered as the angle of s_j with the x-axis. This can be computed using (7)

$$\theta_b^{s_j} = \tan^{-1}[\frac{|y_z - y_z|}{|x_z - x_z|}]\frac{180^o}{\pi}. \tag{7}$$

Next, θ_b is quantized to the closest value of the following eight directions, e.g. $(0^o, 45^o, 90^o, 135^o, 180^o, 225^o, 270^o, 315^o)$ out of which the direction that has majority number of votes, is selected as the dominant direction of b. Finally, the degree of a block (d_b) is computed based on the number of incoming dominant directions from its neighboring blocks. This measure provides an estimate of the activity around that block under consideration from all directions. It is expected that the degree will be higher if an object of interest is present inside the block. This parameter also helps to neutralize errors that might have occurred due to malicious activities such as a person spending more time inside an unimportant block.

3.4 Computation of Block Importance

We describe the computation of the block importance as a 3-step process. The details of these steps are below mentioned.

- **Discarding Rarely Visited Blocks:** In order to compute the importance of blocks in an efficient manner, we first filter out some of the rarely visited blocks, usually considered unimportant, using the global visit count features. First, the minimum and maximum of $g_b \forall b$ are estimated. Then, an approach similar to the average velocity estimation described in the previous section is adopted to construct a histogram. All such blocks where the value of $g_b \forall b$ is less than the average value of the largest bin, are discarded. This essentially discards those blocks where the targets might have visited rarely or not visited at all.
- **Block Popularity Index:** In the next step, we estimate the popularity of a block based on the average instantaneous velocity of the moving targets inside is. This is based on the assumption that a target usually moves slower than its average velocity as it approaches towards an area of interest. Therefore, the instantaneous velocity (v_j^τ) of a target is expected to be lower than its average velocity $(\overline{v}^\tau)$. The popularity index of a block b is computed recursively using the following update equation (8), assuming that the initial popularity index for all blocks is set to zero, i.e. $\rho_b = 0, \forall b \in M$, where ρ_b represents the popularity index of the block b.

$$\rho_b = \rho_b + \frac{\overline{v}^\tau - v_j^\tau}{\overline{v}^\tau} \tag{8}$$

Finally, ρ_b is normalized with global visit count and a metric, we refer to as *"per visit index"* (w_b) is computed using (9). This metric incorporates the importance of each block combining velocity as well as time information.

$$w_b = \forall b \frac{\rho_b}{g_b} \tag{9}$$

– **Computation of Block Importance:** The importance of a block can be computed through independent analysis using of w_b and d_b. However, our systematic experimentation and analysis has revealed that the values of w_b and d_b are highly correlated with each other. Before demonstrating this correlation, we first describe the independent analysis on these variables w_b and d_b, wherein, we convert these estimates into a probability distribution. For example, let w_b be normalized to get the probability of a target being inside block b where its values is taken as $p_w(b)$ and corresponding distribution is given in (10)

$$p_w(b) = P_w(B = b) \quad \text{where } p_w(b) > 0, \forall b \text{ and } \sum_{\forall b} p_w(b) = 1. \tag{10}$$

Similarly, we normalize d_b to get the probability of a target being inside block b and a pdf as described by (11) is assumed.

$$p_d(b) = P_d(B = b) \quad \text{where } p_d(b) > 0, \forall b \text{ and } \sum_{\forall b} p_d(b) = 1 \tag{11}$$

However, the peaks these distributions are likely to be distorted due to measurement noise. In order to adequately suppress such noise and preserve sharp peaks, a zero-phase bi-direction filter has been applied. The zero-phase bi-direction filter is well known for removing noise introduced during feature extraction [17]. Such a filter is usually implemented using a rectangular finite impulse response, $r(.)$ and the filtering operation is performed in both forward and reverse directions as described in equations (13-14)

$$\hat{p}_w(b) = (p_w * r)(b) = \sum_{-\infty}^{+\infty} p_w(m) r(b-m) \tag{12}$$

$$\hat{p}_w^{'} = \text{reverse}(\hat{p}_w)$$

$$\hat{\hat{p}}_w(b) = (\hat{p}_w^{'} * r)(b) = \sum_{-\infty}^{+\infty} \hat{p}_w^{'}(m) r(b-m)$$

$$\dot{p}_w = \text{reverse}(\hat{\hat{p}}_w)$$

$$\hat{p}_d(b) = (p_d * r)(b) = \sum_{-\infty}^{+\infty} p_d(m) r(b-m) \tag{13}$$

$$\hat{p}_d^{'} = \text{reverse}(\hat{p}_d)$$

$$\hat{\hat{p}}_d(b) = (\hat{p}_d^{'} * r)(b) = \sum_{-\infty}^{+\infty} \hat{p}_d^{'}(m) r(b-m)$$

$$\dot{p}_d = \text{reverse}(\hat{\hat{p}}_d).$$

Finally, given a set of real valued discrete time samples, say $p_w(b)$, we get a smoothed signal $\dot{p}_w(b)$ which can be used to detect sharp peaks.

3.5 ROI Localization

We have observed a strong correlation exists between distributions $\dot{p}_w$ and $\dot{p}_d$. Therefore, a simple cross correlation between w_b and d_b can be computed as an evidence. Note that, w_b is estimated on a per-block basis and its value can be quantized into M levels. As, d_b is already represented using 8 discrete levels, we can divide both of these feature spaces into $M \times 8$ subspace and compute the statistics for each interval. Using this sub-space analysis, we demonstrate correlation between w_b and d_b. Finally, the block importance is estimated using (14)

$$p(b = ROI) = \underset{\dot{p}_w \star \dot{p}_d}{\text{argmax}} \tag{14}$$

$$\underset{\dot{p}_w \star \dot{p}_d}{\text{argmax}} \equiv \frac{1}{n-1} \sum_{i=1}^{n} \frac{(\dot{p}_{w,i} - \overline{\dot{p}}_w)}{\sigma_{\dot{p}_w}} \frac{(\dot{p}_{d,i} - \overline{\dot{p}}_d)}{\sigma_{\dot{p}_d}}.$$

A high correlation in a particular interval indicates that both the features agree about the existence of a peak. Therefore, corresponding blocks representing the peak of the distribution can be considered as the location of an interesting object. However, several peaks may be observed in the presence of multiple interesting objects inside a scene.

4 Results and Discussions

In this section, we discuss various systematic experiments carried out to evaluate our proposed algorithm and benchmark it against state-of-the-art baseline methods using three publicly available surveillance datasets and a scenario specific in-house dataset.

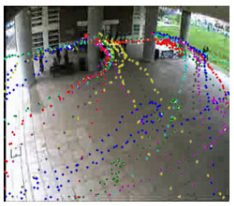
4.1 Datasets and Baselines

Four datasets have been been considered for experimental validation. Three of them, namely, CAVIAR [5], ViSOR[1], and Grand Central Station Dataset (CUHK) [35] are well known public datasets often used by the surveillance research community. In addition, a set of scenario-specific in-house videos were also used during evaluation. Although those datasets are not recorded for region of interest,We have found some clips to test our hypothesis; we search and picked video clips of 240 seconds duration from the CAVIAR dataset. In these clips, 2 or 3 moving targets can be seen moving freely. These targets interact with interesting static objects such as a vending machine and a bookshelf. In contrast, the ViSOR dataset contains outdoor videos [28]. These videos are long sequences, typically of 40-60 minutes in duration. Grand central station dataset or CUHK dataset [35] was recorded inside an underground station. It is a 34 minutes long video with nearly 700 number of curtailed trajectories. We have merged these curtailed trajectories and created a set of 40 clean trajectories. In addition to above three datasets, we have tracked human movements inside a laboratory environment, which we refer as the In-HOUSE dataset.

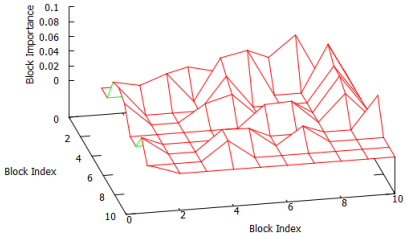
ROI in respect of a scene may be different due to the objective of the Interest. Last few years a handfull of method prposed by so many researchers with a similar goal; to identify region of interest in a video sequence. Region of Interest may be in image level, like object localization and moving object localization; or may be in video level(i.e image and motion mixed) like, scene classification, Abandoned Object localization etc. We present comparisons against relevant baseline algorithms having a similar objective of e.g. ROI localization. Image-guided ROI localization is primarily based on extracting salience locations or visual attention features that indicate the presence of a ROI. In contrast, video-based techniques use spatio-temporal correlation to detect interesting locations. For comparison with image-guided techniques, we have selected salience-map based interest area localization method discussed in [7, 18, 23], and a saliency-combined visual attention based model presented in [1]. It is to be noted that, above methods do not consider the temporal aspects. They are primarily designed using spatio-visual features. For example, Rahtu et al. [23] have proposed an image and video object segmentation technique that combines salience with a conditional random field (CRF) model to localize interest areas. The DRFI based method proposed by Jiang et al. [7] has been applied on several video frames of the selected datasets and a comparison has been presented. A similar approach has been adopted while comparing the results against the method proposed in [18] and [1]. In order to compare with existing video-guided methods , we have used the abandoned object detection[2] and the trajectory clustering based method proposed by Bharath et al. [1]. We present comparative results with above state-of-the-art techniques

[1] `http://www.openvisor.org`

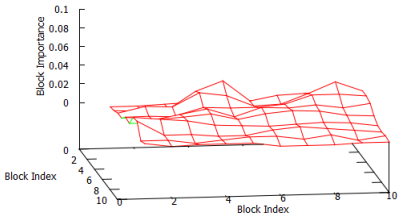
[2] http://www.mathworks.in/help/vision/examples/abandoned-object-detection.html.



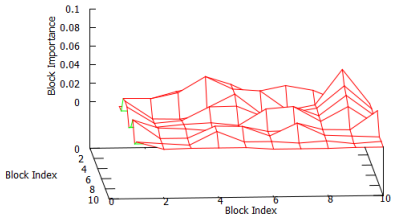
5.1 Surveillance scene with superimposed target trajectories.



5.2 Localization of the ROI using w_b.



5.3 Localization of the ROI using d_b.



5.4 Localization of areas using a combination of w_b and d_b.

5.5 Localized ROI using the proposed methodology.

Fig. 5 Localization of ROI using the proposed method applied a sample video clip from the ViSOR dataset.

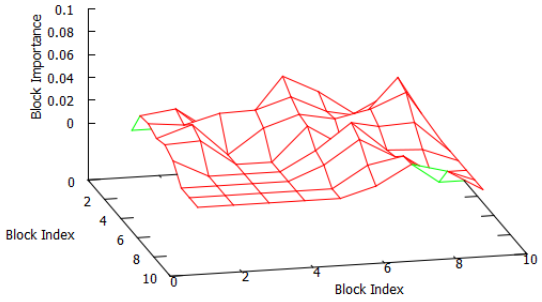
to demonstrate the superiority of the proposed algorithm. A summery on those algorithm can be found in Table 1.

4.2 Experimental Results using Various Datasets

To begin with, we have applied w_b and d_b independently for localizing ROI on the scenes from CAVIAR, ViSOR, CUHK, and In-HOUSE datasets. In the next phase, a combination of w_b and d_b as described in section 3, is used for localization. Figure 5 presents results of localization in ViSOR scene. It is evident from Figure 5.5 that, the inclusion of direction into block importance computation has significantly improved the accuracy of localization as compared to velocity based localization. We extend a similar analysis using

ROI Method	Input	Application	ROI
[6]	Video as a Sequence of Frames	Abandoned object detection.	Abandoned Object
[7]	Frames	Object Segment	Object Present in Scene
[23]	Image Sequence and Motion Information	Image and motion based ROI detection.	Moving Object Segmentation
[1]	Video as Frame Sequence	Understand Scene by identifying Object Locaton	Moving Object Location
[18]	Frame as Image	Unique Image Patch Identification	Objects in Image
Proposed	Scene as a Video	Long term activity based ROI detection	Interest Region of Moving Objects

Table 1 ROI detection Methods



6.1 Localization of areas of interest using a combination of w_b and d_b on CAVIAR videos.

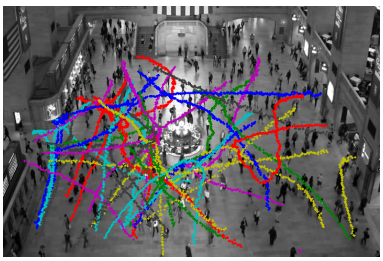
6.2 Identification of ROI (reading desk) on the scene.

6.3 Identification of ROI (ATM machine) on the scene.

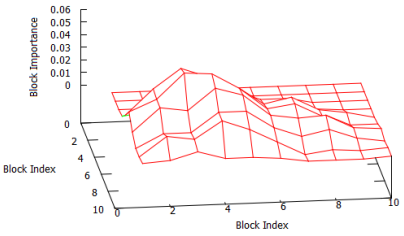
Fig. 6 Localization of interest areas using the proposed method applied on videos of CAVIAR dataset.

the videos taken from CAVIAR, CUHK, and In-HOUSE datasets. Results of such analysis are presented in Figures 6-8. As depicted in Figures 6.2-6.3, the scene representing CAVIAR contains two ROI, e.g. book shelf, and vending machine. The proposed algorithm was able to localize both. Presence of peaks over these regions proves our claim (refer to Figure 6.1).

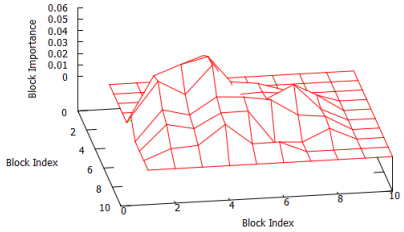
The results on the CUHK dataset are presented in Figure 7. It may be observed from Figure 7.1 that the trajectory density around the central portion of the scene is high as compared to the surrounding areas. This is because, a large number of targets visited the central lounge area. We have divided the scene into 10×10 blocks and the variations of importance across these blocks are presented in Figures 7.2-7.4. We have also verified that, though both features perform consistently, the combined feature set performs more accurately to localize the ROI in comparison to the ground truth.



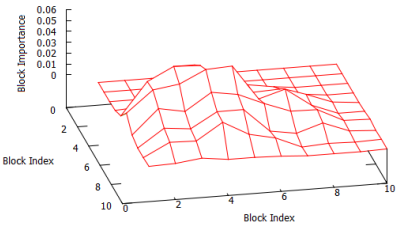
7.1 Background of the CUHK grand central dataset and overlaying of 40 clean trajectories constructed from 700 curtailed trajectories.



7.2 Localization of areas of interest using d_b.



7.3 Localization of areas of interest using w_b.



7.4 Localization of areas of interest using a combination of w_b and d_b.

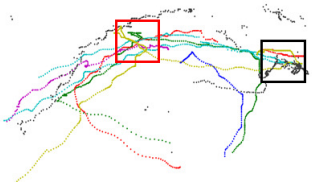
Fig. 7 Localization of ROI using CUHK dataset videos.

The results of ROI localization (e.g. center table) using In-HOUSE dataset are presented in Figure 8. It is clear that the proposed algorithm successfully localizes the center table while rejecting a similar high density areas (marked with rectangular box of black boundary).

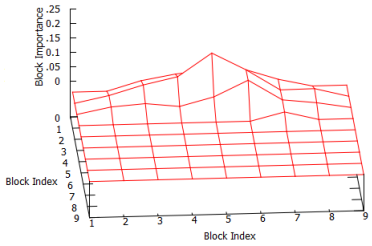
In addition to the qualitative analysis as above, we have also performed a visual analysis on target trajectories to support the results already obtained. For example, the presence of a probable ROI can be verified from the plots of trajectories (x-axis represents the cumulative frame number and y-axis represents block number) as shown in Figure 9.1 and we have highlighted relevant time-sequences that correspond to locations of the ROI. Localizing such segments through time-series analysis may not be trivial since such segments may also appear due to several other reasons. However, our proposed algorithm was successful in correctly localizing those blocks where targets spent time due to the presence of other static objects of interest. A few snapshots revealing targets visiting the centre table in the IN-HOUSE dataset are shown in Figure 9.2.

4.3 Effect of Model Parameters

It has been observed that our proposed algorithm is sensitive to block size. We have carried out analysis by varying block size to understand its impact on



8.1 Trajectory density and corresponding locations denoting probable ROI.

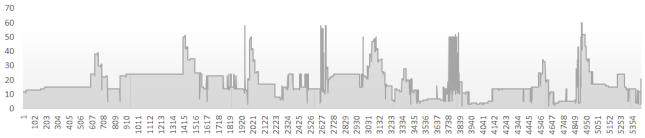


8.2 Peak representing the area of the centre table that was found using w_b and d_b in combination.

8.3 Localization of a probable ROI of In-HOUSE surveillance scene.

8.4 Location of an area that was detected as ROI using w_b feature only, whereas the same was rejected when w_b and d_b were used in combination.

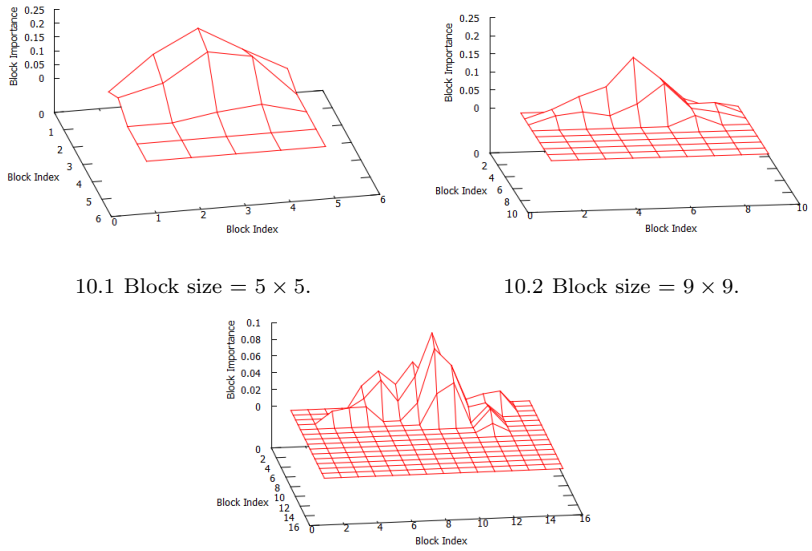
Fig. 8 Localization of ROI using the proposed method applied on videos of In-HOUSE dataset.



9.1 Temporal localization of the ROI (center table) from the cumulative trajectory constructed by concatenating target trajectories of In-HOUSE dataset.

9.2 Some of the key frames extracted from the videos when users are passing through the center table.

Fig. 9 Localization of the ROI, e.g. center table, in videos of In-HOUSE dataset.



10.1 Block size = 5×5.

10.2 Block size = 9×9.

10.3 Block size = 15×15.

Fig. 10 Localization of ROI in videos of In-HOUSE dataset with varying block sizes.

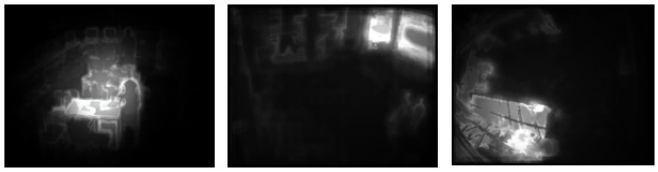
the accuracy of localization. The distribution of block importance for varying block sizes, e.g. 5×5, 9×9, and 15×15, are presented in Figure 10. It is evident that the peaks representing probable ROI become more obvious as block size reduces. A larger block size effectively reduces the total number of inter-block movements, thus resulting in peaks with larger variance. However, if the block size is reduced beyond a scene-specific threshold, more dense peaks may be observed. This will essentially make the localization inaccurate. After experimental cross validation, a grid size of 9×9 has been found to be optimal for the datasets used in our experiments.

4.4 Comparative Performance Analysis

We have compared our results against popular baselines mentioned earlier. In Figures 11.2-11.4, results using baseline techniques are presented. It may be observed that, salience based techniques [7, 18, 23] fail to identify the ROIs in CAVIAR dataset.

Results using baseline techniques applied on videos are presented in Figure 12.1-12.3. After carefully analyzing the results obtained using the baseline algorithms, it is possible to conclude that our proposed algorithm was more successful in detecting ROI while rejecting the false positives (e.g. locations with high trajectory density), as against the chosen baseline techniques.

11.1 Scenes of various datasets used in comparisons.



11.2 Technique proposed by Rathu et al. [23].

11.3 Technique proposed by Jiang et al. [7].

11.4 Salience guided interest point segmentation proposed in [18].

Fig. 11 Comparative performance analysis of the proposed ROI localization algorithm against popular image-guided interest area localization techniques.

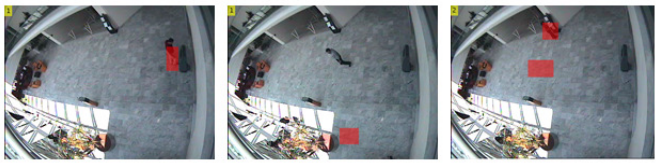
4.5 Verification of the Theoretical Model

We have verified the results obtained using our proposed methodology with the theoretical model described in section 3.1 using Kullback-Leibler Divergence (KLD). KLD can be computed for a pair of probability distributions using (15), where $\dot{p}(I(b))$ and $p^T(I(b))$ represent probability distribution of the importance of blocks and the theoretical formulation in section 3.1, respectively.

$$D_{\mathrm{KL}}(\ p^T(I(b))\ ||\ \dot{p}(I(b))\ ) = \sum_y \ln\left(\frac{p^T(I(b))}{\dot{p}(I(b))}\right) p^T(I(b)) \tag{15}$$

The quantity is often used for feature selection in classification problems, where $P(y)$ and $\dot{p}(I(b))$ represent conditional distributions of the feature under two different classes. To verify the hypothesis, the distribution shown in

12.1 Abandoned object detection in videos proposed in [6].



12.2 Trajectory density based interest point localization.

12.3 Method proposed by Bharath et al. [1].

Fig. 12 Comparative performance analysis of the proposed ROI localization algorithm against popular video-guided interest area localization techniques.

Table 2 Divergence values using varying grid configurations and combination of features.

Feature	KL Divergence(D_{KL}) Varying Grid Configurations			
	5×5	9×9	15×15	20×20
w_b	0.6956	0.5497	0.9619	0.8668
w_d	0.7863	0.6142	0.9128	0.9347
w_b and w_d	0.5195	0.4872	0.8725	0.8208

Figure 3.9 was taken as $p^T(I(b))$ and the distributions shown in Figure 10 were considered as $\dot{p}(I(b))$. We computed divergence values using varying grid sizes with independent as well as combined features and the measured values are presented in Table 2.

The smaller the coefficient, the more accurate the matching. Therefore, our analysis confirms that the results of our algorithm is in agreement with the theoretical model when the scene is divided into 9×9 blocks. It can be observed that, performance improves significantly when metric of importance is computed jointly using w_b and w_d as against independent features. This has also been verified through correlation coefficients: $p^T(I(b))$ and $\dot{p}(I(b))$. Correlation coefficients with varying block sizes are presented in Table 3.

It may be noted from the above results that both metrics unanimously agree upon 9×9 block size to be most suitable choice for the present analysis.

Table 3 Correlation coefficient(ρ) values using varying grid configurations and combination of features.

Feature	Correlation coefficient(ρ)			
	Varying Grid Configurations			
	5×5	9×9	15×15	20×20
w_b	0.01	0.77	0.63	0.37
w_d	0.21	0.65	0.52	0.19
w_b and w_d	0.17	0.81	0.67	0.29

5 Conclusion and Future Work

In this paper, a technique for localizing ROI by analyzing motion trajectories of moving targets, has been proposed. The proposed method is based on maximizing the correlation of motion dynamics features. A theoretical assumption about the natural target motion inside as unconstrained environment, has been proposed and further statistically validated using various publicly available video surveillance datasets. The results of our experiments demonstrate the ability of the proposed methodology to localize key areas in a given scene. It is anticipated that, the proposed work has good potential to throw insight into human behavior understanding in the context of visual surveillance.

Several extension of the present work are possible. For example, the scene can be labeled based on the importance value of the blocks and the background can be segmented into meaningful regions. Feature describing these local regions can be extracted and they can be used in the decision making process. We also plan to introduce a Bayesian framework where likelihood and prior distribution parameters can be estimated from ground truth data. That is, Equations (8), (9) and (10) can be modified accordingly. Additionally, the posterior probability ($p_w(b)$) can be normalized using the summed evidence (equivalent to the function of term $p_w(b)$). A similar formulation can also be used for posterior calculation of $p_d(b)$. It is anticipated that, such extensions will further strengthen the principle of our algorithm.

References

1. R. Bharath, L. Nicholas, and X. Cheng. Scalable scene understanding using saliency-guided object localization. In *Control and Automation, Proceedings of the IEEE International Conference on*, pages 1503–1508, 2013.
2. L. Brun, A. Saggese, and M. Vento. Dynamic scene understanding for behavior analysis based on string kernels. *Circuits and Systems for Video Technology, IEEE Transactions on*, 24(10):1669–1681, Oct 2014.
3. T. Dinh, N. Vo, and G. Medioni. Context tracker: Exploring supporters and distracters in unconstrained environments. In *Computer Vision and Pattern Recognition, Proceedings of the IEEE Computer Society Conference on*, pages 1177–1184, June 2011.
4. D. Dogra, A. Ahmed, and H. Bhaskar. Interest area localization using trajectory analysis in surveillance scenes. In *10th International Conference on Computer Vision Theory and Applications, Proceedings of the*, pages 478–485, March 2015.

5. R. Fisher, J. Santos-Victor, and J. Crowley. Caviar: Context aware vision using image-based active recognition. `http://homepages.inf.ed.ac.uk/rbf/CAVIAR/`, 2001. Accessed: July 2014.
6. Mathworks Inc. Abandoned object detection. `http://www.mathworks.in/help/vision/examples/abandoned-object-detection.html`, 2014. Accessed: July 2014.
7. H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *Computer Vision and Pattern Recognition, Proceedings of the IEEE Computer Society Conference on*, pages 2083–2090, 2013.
8. P. Kapsalas, K. Rapantzikos, A. Sofou, and Y. Avrithis. Regions of interest for accurate object detection. In *Content-Based Multimedia Indexing, Proceedings of the International Workshop on*, pages 147–154, June 2008.
9. J. Keum, H. Lee, and M. Hagiwara. Mean shift-based sift keypoint filtering for region-of-interest determination. In *Soft Computing and Intelligent Systems and International Symposium on Advanced Intelligent Systems, Proceedings of the International Conference on*, pages 266–271, Nov 2012.
10. G. Kim and A. Torralba. Unsupervised detection of regions of interest using iterative link analysis. In *Advances in Neural Information Processing Systems, Proceedings of the*, pages 961–969, 2009.
11. Y. Lai and C. Yang. Video object retrieval by trajectory and appearance. *Circuits and Systems for Video Technology, IEEE Transactions on*, 25(6):1026–1037, June 2015.
12. Wen-Fu Lee, Tai-Hsiang Huang, Su-Ling Yeh, and Homer H Chen. Learning-based prediction of visual attention for video signals. *Image Processing, IEEE Transactions on*, 20(11):3028–3038, 2011.
13. Jia Li, Yonghong Tian, Tiejun Huang, and Wen Gao. Probabilistic multi-task learning for visual saliency estimation in video. *International journal of computer vision*, 90(2):150–165, 2010.
14. W. Lin, Y. Zhang, J. Lu, B. Zhou, J. Wang, and Y. Zhou. Summarizing surveillance videos with local-patch-learning-based abnormality detection, blob sequence optimization, and type-based synopsis. *Neurocomputing*, 155(0):84 – 98, 2015.
15. Tie Liu, Nanning Zheng, Wei Ding, and Zejian Yuan. Video attention: Learning to detect a salient object sequence. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.
16. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *Computer Vision and Pattern Recognition, Proceedings of the IEEE Computer Society Conference on*, pages 1988–1995, June 2009.
17. M. Manikandan and K. Soman. A novel method for detecting r-peaks in electrocardiogram (ecg) signal. *Biomedical Signal Processing and Control*, 7(2):118–128, 2012.
18. R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *Computer Vision and Pattern Recognition, Proceedings of the IEEE Computer Society Conference on*, pages 1139–1146, June 2013.
19. S. Mitri, S. Frintrop, K. Pervolz, H. Surmann, and A. Nuchter. Robust object detection at regions of interest with an application in ball recognition. In *Robotics and Automation, Proceedings of the IEEE International Conference on*, pages 125–130, April 2005.
20. B. Morris and M. Trivedi. Learning and classification of trajectories in dynamic scenes: A general framework for live video analysis. In *Advanced Video and Signal Based Surveillance, Proceedings of the IEEE Fifth International Conference on*, pages 154–161, Sept 2008.
21. W. Osberger and A. Rohaly. Automatic detection of regions of interest in complex video sequences. In *Photonics West-Electronic Imaging, Proceedings of the*, pages 361–372, 2001.
22. C. Piciarelli, C. Micheloni, and G. Foresti. Trajectory-based anomalous event detection. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1544–1554, Nov 2008.
23. E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä. Segmenting salient objects from images and videos. In *European Conference on Computer Vision, Proceedings of the*, pages 366–379. Springer, 2010.

24. M. Rokunuzzaman, K. Sekiyama, and T. Fukuda. Automatic roi detection and evaluation in video sequences based on human interest. *Journal of Robotics and Mechatronics*, 22(1):65–75, 2010.
25. I. Saleemi, K. Shafique, and M. Shah. Probabilistic modeling of scene dynamics for applications in visual surveillance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(8):1472–1485, 2009.
26. N. Shou, H. Peng, H. Wang, L. Meng, and K. Du. An rois based pedestrian detection system for single images. In *Image and Signal Processing, Proceedings of the International Congress on*, pages 1205–1208, Oct 2012.
27. N. Suzuki, K. Hirasawa, K. Tanaka, Y. Kobayashi, Y. Sato, and Y. Fujino. Learning motion patterns and anomaly detection by human trajectory analysis. In *Systems, Man and Cybernetics, Proceedings of the IEEE International Conference on*, pages 498–503, Oct 2007.
28. R. Vezzani and R. Cucchiara. Video surveillance online repository (visor): an integrated framework. *Multimedia Tools and Applications*, 50(2):359–380, 2010.
29. W. Wang, W. Lin, Y. Chen, J. Wu, J. Wang, and B. Sheng. Finding coherent motions and semantic regions in crowd scenes: A diffusion and clustering approach. In *European Conference on Computer Vision, Proceedings of the*, volume 8689 of *Lecture Notes in Computer Science*, pages 756–771, 2014.
30. X. Wang, K. Tieu, and E. Grimson. Learning semantic scene models by trajectory analysis. In *European Conference on Computer Vision, Proceedings of the*, pages 110–123. Springer, 2006.
31. M. Xiang, F. Bashir, A. Khokhar, and D. Schonfeld. Event analysis based on multiple interactive motion trajectories. *Circuits and Systems for Video Technology, IEEE Transactions on*, 19(3):397–406, March 2009.
32. D. Xu, X. Wu, D. Song, N. Li, and Y. Chen. Hierarchical activity discovery within spatio-temporal context for video anomaly detection. In *Image Processing, Proceedings of the IEEE International Conference on*, pages 3597–3601, Sept 2013.
33. M. Xuan, V. Monga, R. Bala, and F. Zhigang. Adaptive sparse representations for video anomaly detection. *Circuits and Systems for Video Technology, IEEE Transactions on*, 24(4):631–645, April 2014.
34. Yun Zhai and Mubarak Shah. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 815–824. ACM, 2006.
35. B. Zhou, X. Wang, and X. Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *Computer Vision and Pattern Recognition, Proceedings of the IEEE Computer Society Conference on*, pages 2871–2878, 2012.
36. Y. Zhou, S. Yan, and T. Huang. Detecting anomaly in videos from trajectory similarity analysis. In *Multimedia and Expo, Proceedings of the IEEE International Conference on*, pages 1087–1090, July 2007.

Sk. Arif Ahmed has obtained a master degree in Computer Applications form West Bengal University of Technology. Currently working as an Assistant Professor at Haldia institute of Technology, India. His area of interest is in the domain of computer vision, image processing, and scene analysis.

Dr. Debi Prosad Dogra is an Assistant Professor in the School of Electrical Sciences, IIT Bhubaneswar, India. Prior to joining, IIT Bhubaneswar, Dr. Dogra was with Advanced Technology Group, Samsung Research Institute Noida, India for a period of two years (2011-2013). In SRI Noida, Dr. Dogra was leading a research team, mainly doctorates where his main area of focus was in designing applications in the domains of healthcare automation, gesture recognition, augmented reality with the help of video object tracking and image segmentation, visual surveillance. Prior to joining SRI Noida, he obtained his Ph.D. degree from IIT Kharagpur in the year of 2012. He received his M.Tech degree from IIT Kanpur in 2003 after completing his B.Tech. (2001) from HIT Haldia, India. After finishing his masters, he joined Haldia Institute of Technology as a faculty members in the Department of Computer Sc. & Engineering (2003-2006). He has worked with ETRI, South Korea during 2006-2007 as a researcher. Dr. Dogra has published more than 15 international journal and conference papers in the areas of computer vision, image segmentation, and healthcare analysis. He is a member of IEEE.

Dr. Byung-Gyu Kim has received his BS degree from Pusan National University, Korea, in 1996 and an MS degree from Korea Advanced Institute of Science and Technology (KAIST) in 1998. In 2004, he received a PhD degree in the Department of Electrical Engineering and Computer Science from Korea Advanced Institute of Science and Technology (KAIST).
In March 2004, he joined in the real-time multimedia research team at the Electronics and Telecommunications Research Institute (ETRI), Korea where he was a senior researcher. In ETRI, he developed so many real-time video signal processing algorithms and patents and received the Best Paper Award in 2007.
In February 2009, he joined the Division of Computer Science and Engineering at SunMoon University, Korea where he is currently a professor. In 2007, he served as an editorial board member of the International Journal of Soft Computing, Recent Patents on Signal Processing, Research Journal of Information Technology, Journal of Convergence Information Technology, and Journal of Engineering and Applied Sciences. Also, he is serving as an associate editor of Circuits, Systems and Signal Processing (Springer), The Journal of Supercomputing (Springer), The Journal of Real-Time Image Processing (Springer), The Scientific World Journal (Hindawi), and International Journal of Image Processing and Visual Communication (IJIPVC). He also served as Organizing Committee of CSIP 2011 and Program Committee Members of many international conferences. He has received the Special Merit Award for Outstanding Paper from the IEEE Consumer Electronics Society, at IEEE ICCE 2012, Certification Appreciation Award from the SPIE Optical Engineering in 2013, and the Best Academic Award from the CIS in 2014. He has been honored as an IEEE Senior member in 2015. He has published over 130 international journal and conference papers, patents in his field. His research interests include software-based image and video object segmentation for the content-based image coding, video coding techniques, 3D video signal processing, wireless multimedia sensor network, embedded multimedia communication, and intelligent information system for image signal processing. He is a senior member of IEEE and a professional member of ACM, and IEICE.

Dr. P. Hill is a senior research teacher in faculty of Engineering in University of Bristol. He is also a research fellow in Image Communication. He has published more than 35 papers in international journals and conferences. His research interest includes image fusion, image segmentation, texture analysis, image compression. He has obtained B.Sc (open), M.Sc. and Ph.D.

Dr. Harish Bhaskar is an Assistant Professor in the Department of Electrical and Computer Engineering at Khalifa University (KUSTAR), Abu Dhabi, U.A.E. Dr. Bhaskar also currently holds a honorary researcher position at the University of Bristol, U.K. Prior to rejoining KUSTAR earlier in 2014, Dr. Bhaskar spent nearly a year as a Chief Engineer at Samsung Electronics, Noida, INDIA within the Advanced Software Group, Systems Team. Dr. Bhaskar's transit to the industry came after completing 4 years as an Assistant Professor of Computer Engineering at KUSTAR, since 2009. During these years, Dr. Bhaskar has been actively involved in teaching and research within computer vision and image processing. In addition, Dr. Bhaskar also plays a key role in the Visual Signal Analysis and Processing (VSAP) research center that has been jointly established between KUSTAR and the University of Bristol, U.K. Before moving to the U.A.E. for an academic career, Dr. Bhaskar worked as a post-doctoral researcher at both the University of Manchester and Lancaster University; U.K. Dr. Bhaskar has been actively associated with several European research institutes and the Ministry of Defense U.K. His primary research interests are in the field of computer vision, image processing, data mining, visual cryptography, medical imaging and robotics.